統計學習初論

第三組

黃啟宏 B06705002

許亦佑 B06705021

吳禹辰 B06705034

王松億 B06705049

Frame the problem

在近幾年，信用卡債務危機給銀行帶來的麻煩著實令銀行管理者頭痛不已。不論使用者的付費能力如何。使用者常常會過量地使用信用卡。當然、有償還能力的使用者是沒什麼問題的，但是那些沒有償還能力的使用者將會有拖欠付款的行為。因此，若是對所有的使用者都給予相同的額度，後者若是在過量的使用信用卡後卻無法在時效內付款，將會導致銀行的資金出現問題。我們想要針對這個問題，利用信用卡使用者的基本資料和使用者的消費取向，例如性別、每月使用金額、教育程度、婚姻狀態...等等，分析信用卡使用者在**下個月**是否會拖欠，進而讓銀行可以判斷是否給予不同的使用者不同的信用卡額度。

在查閱資料後和教授的說明後，我們發現我們的想法和2009年的一篇論文不謀而合。因此，我們根據這篇論文的研究方法作為基礎，進一步實作其餘資料處理和分析資料的方式，試圖做出更甚於之的研究成果。

Get the data

Default Payments of Credit Card Clients in Taiwan from 2005

[Default of Credit Card Clients Dataset]

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

Explore the data

Field name	Description
ID	ID of each client
LIMIT_BAL	Amount of given credit in NT dollars (includes individual and family/supplementary credit)
SEX	Gender (1=male, 2=female)
EDUCATION	(1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY_0	Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay f or two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
PAY_2	Repayment status in August, 2005 (scale same as above)
PAY_3	Repayment status in July, 2005 (scale same as above)
PAY_4	Repayment status in June, 2005 (scale same as above)
PAY_5	Repayment status in May, 2005 (scale same as above)
PAY_6	Repayment status in April, 2005 (scale same as above)
BILL_AMT1	Amount of bill statement in September, 2005 (NT dollar)
BILL_AMT2	Amount of bill statement in August, 2005 (NT dollar)
BILL_AMT3	Amount of bill statement in July, 2005 (NT dollar)
BILL_AMT4	Amount of bill statement in June, 2005 (NT dollar)
BILL_AMT5	Amount of bill statement in May, 2005 (NT dollar)
BILL_AMT6	Amount of bill statement in April, 2005 (NT dollar)
PAY_AMT1	Amount of previous payment in September, 2005 (NT dollar)
PAY_AMT2	Amount of previous payment in August, 2005 (NT dollar)
PAY_AMT3	Amount of previous payment in July, 2005 (NT dollar)
PAY_AMT4	Amount of previous payment in June, 2005 (NT dollar)
PAY_AMT5	Amount of previous payment in May, 2005 (NT dollar)
PAY_AMT6	Amount of previous payment in April, 2005 (NT dollar)
default.payment.next.month	Default payment (1=yes, 0=no)

在此資料集中，總共有30000筆資料。

因為我們想要分析的是使用者在下個月是否會拖欠，因此我們將default payment next month設為標籤，並將其餘所有欄位設成特徵。

因為我們希望可以透過改量論文中的過程以取得最佳的結果，因此在資料分群上我們照著論文中的數量比將訓練資料和測試資料的數量設定為1比1。

Prepare the data to better expose the underlying data patterns to Machine Learning algorithms

在研究完論文後，首先，我們認為論文中對於資料的處理上有不足。在原先的資料中，所有的特徵都沒有被處理，而是直接照著原數據放入模型中訓練。我們認為這些資料可以被更加妥善地處理。

首先，我們發現：SEX、EDUCATION和MARRIAGE這三項特徵都屬於類別而不是屬於連續變數，因此我們認為，對其取虛擬變數能夠藉此檢驗不同屬性類型對因變數的作用，提高模型的精準度。

接著，我們對整個資料集的特徵進行標準化，希望每個特徵值可以對結果做出相近程度的貢獻，藉此提高模型的準確率。

在這之後，我們希望除了論文已有的模型，增加一些至今為此通過這堂課所學習過的回歸模型，比較各種模型之間的差異和適合度。

Explore many different models and short-list the best ones

我們根據論文中的各種模型分別進行model training，並根據論文，比較以下四種論文中提到的指標來檢視預測準確度而得到以下結果。

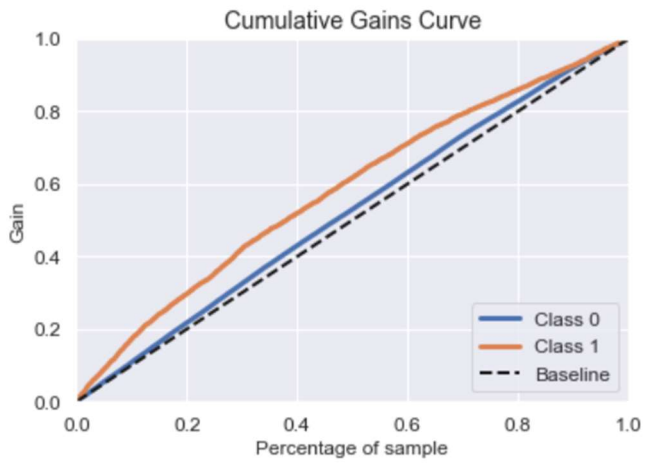
1) Cumulative Gain Curve

X軸為test data數量的比例，Y軸則為經過模型挑選後達到目標的比例。

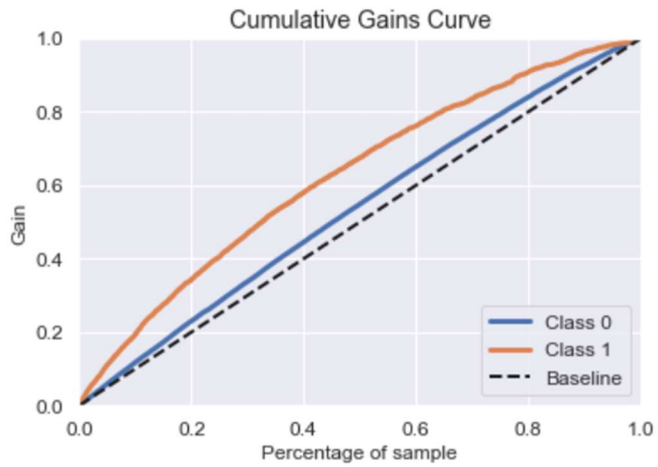
此圖有三條曲線，分別為Baseline、Class 0和Class 1。

Baseline代表的是沒有採用模型，隨機從群體挑選出此百分比的人之中，結果的class為X的人佔整個群體class為X的比例，因此為斜率為1之直線(隨機)。Class 0和Class 1則各代表著採用此模型後各百分比之下的目標達成率。

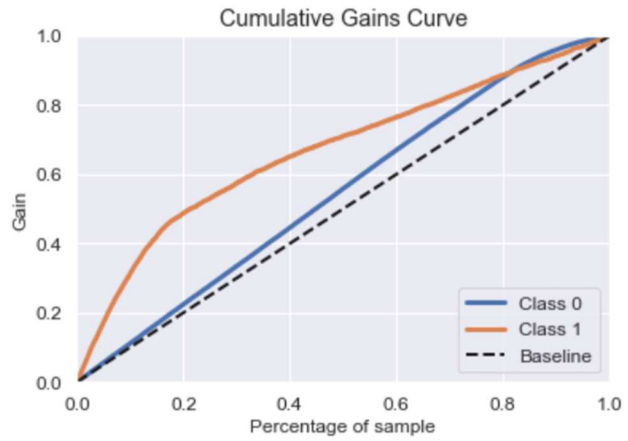
KNN



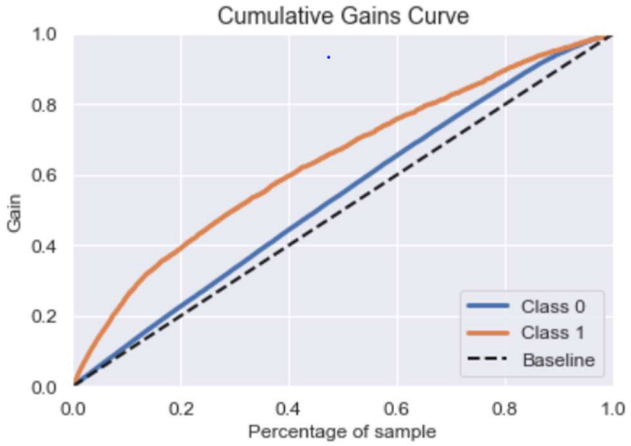
Logistic



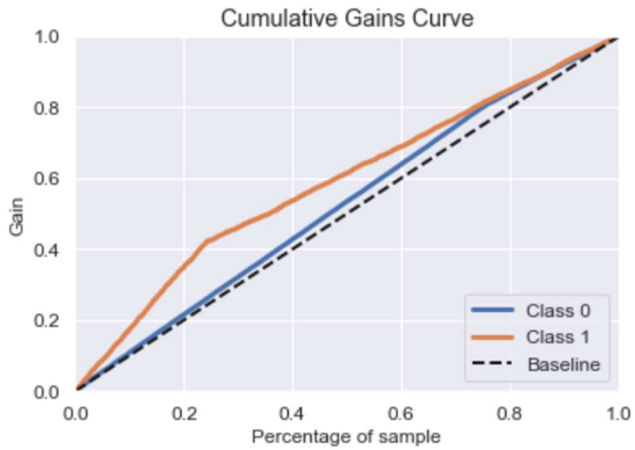
Discriminant Analysis



Naive Bayesian Classifier



Decision Tree



Class 1和class 0離Baseline越遠的模型，將會是越好的模型(可以更精準的定位目標)。

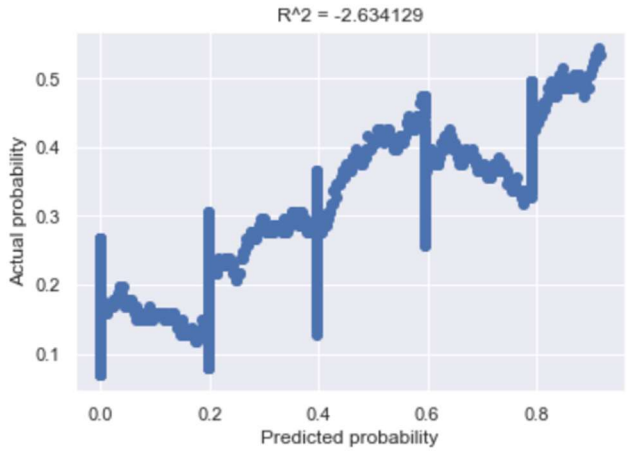
我們可以從上圖知道，在此指標中，Discriminant Analysis表現較好。

2) Sorting Smoothing Method(SSM)

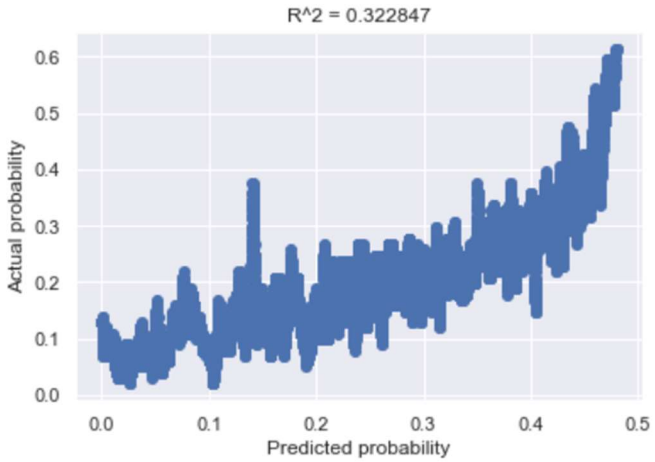
X軸為預測會拖欠的概率，Y軸為估計的真實拖欠概率(利用SSM算出)。

首先，我們在利用各種模型預測會拖欠的概率後，以此為基準對這些資料進行從小到大的排序。接著，對每筆資料的真實數據前後各50筆取平均值，作為此筆資料的真實拖欠概率，作為判斷比較模型好壞的基準。

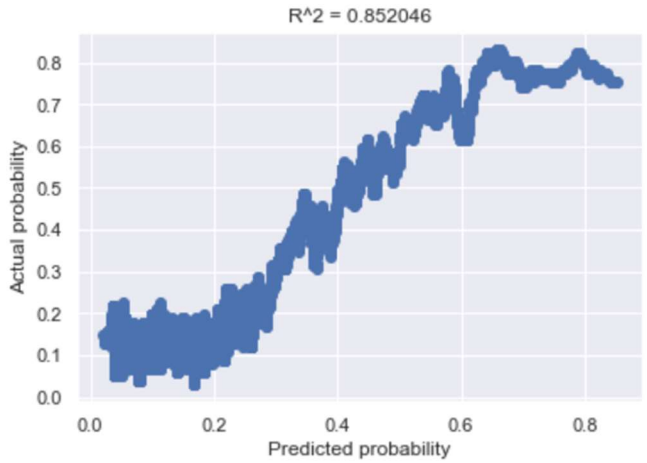
KNN



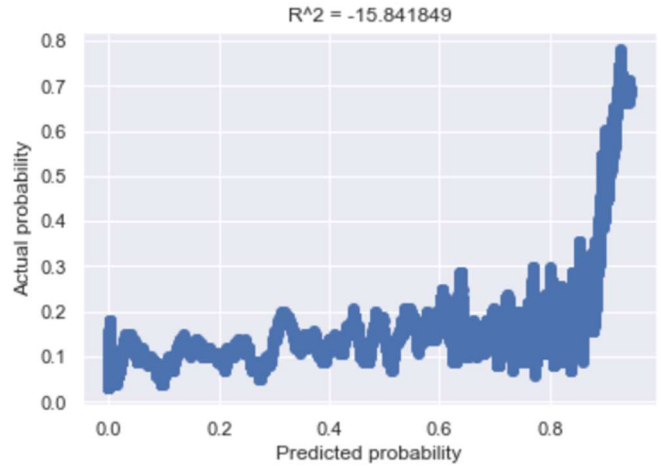
Logistic



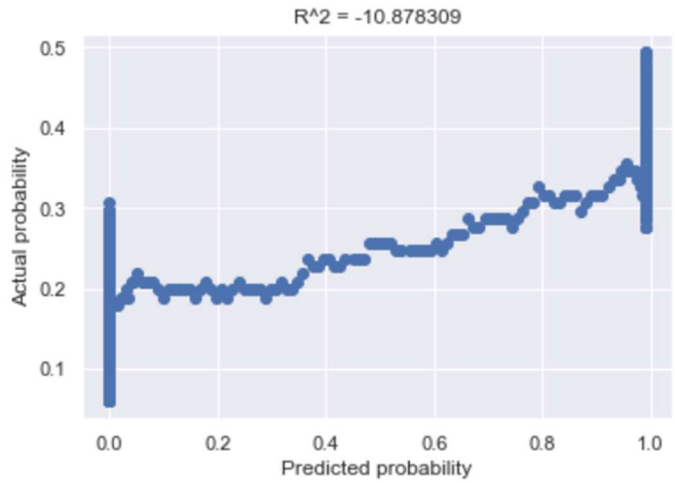
Discriminant Analysis



Naive Bayesian Classifier



Decision Tree



R^2為解釋力，當R^2越高，且當預測拖欠概率越接近於估計真實拖欠概率時(斜率為1)，我們判斷此模型的精確度較高。從上圖可以得知，Discriminant Analysis的解釋力較高，且預測拖欠概率越相當接近於估計真實拖欠概率，因此判斷此模型精確度較高。

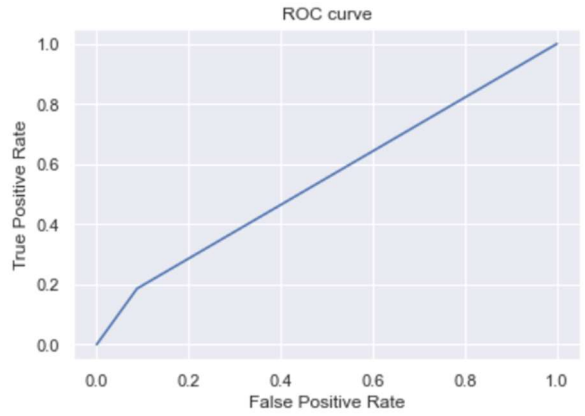
3) ROC curve

X軸為偽陽性的概率，Y軸為真陽性的概率。

ROC是坐標圖式的分析工具，離左上角越近的點預測準確率越高。AUC為ROC曲線下方的面積，為正確判斷陽性樣本的值高於陰性樣本之機率 (AUC值越大，正確率越高)。

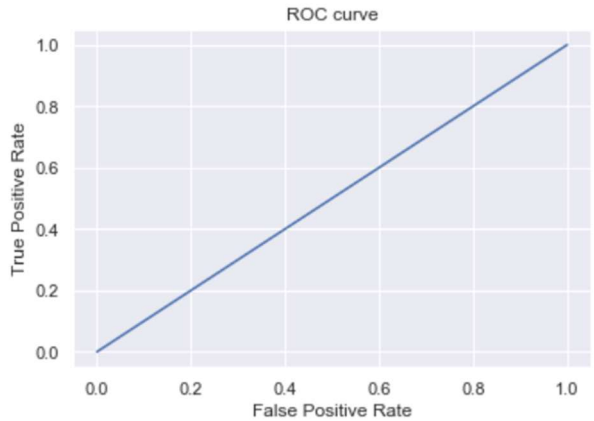
KNN

```
Area ratio: 0.5491262220601177
```



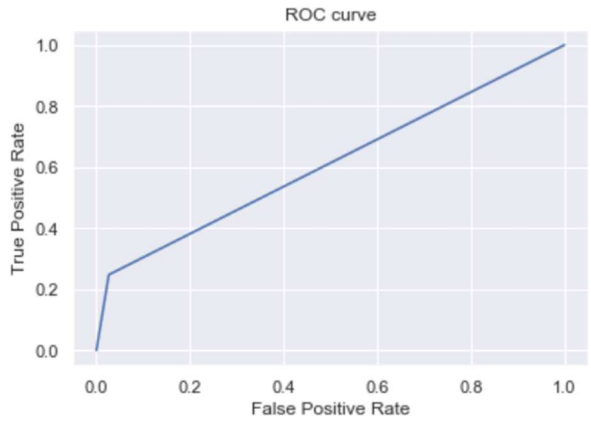
Logistic

```
Area ratio: 0.5001529987760098
```



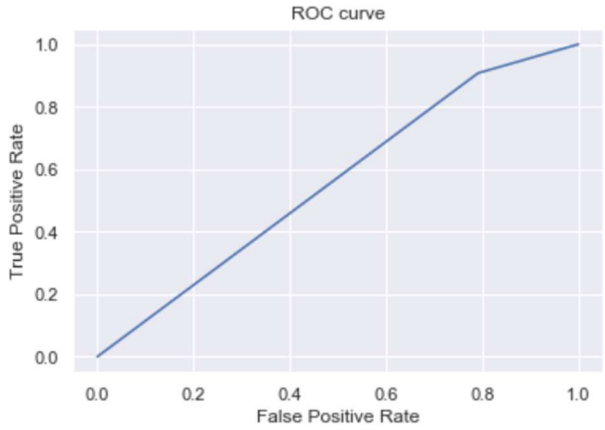
Discriminant Analysis

```
Area ratio: 0.6105216627070256
```



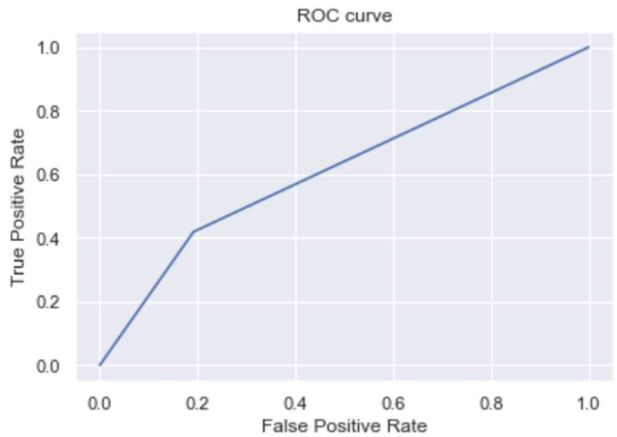
Naive Bayesian Classifier

```
Area ratio: 0.558081788669932
```



Decision Tree

```
Area ratio: 0.6138019815036843
```



我們利用AUC大小比較模型的好壞。從上圖可以得知，Discriminant Analysis和Decision Tree的AUC較高，模型表現較好。

4) Error Rate

利用模型預測出拖欠機率，若是大於0.5則視為1，小於0.5視為0。接著，和真實數據做對比，計算錯誤率。

model	error rate
KNN	0.246
Logistic	0.218
Discriminant Analysis	0.185
Naive Bayesian Classifier	0.639
Decision Tree	0.276

從上述結果，我們可以得知Logistic和Discriminant Analysis模型表現的比較好。

結論

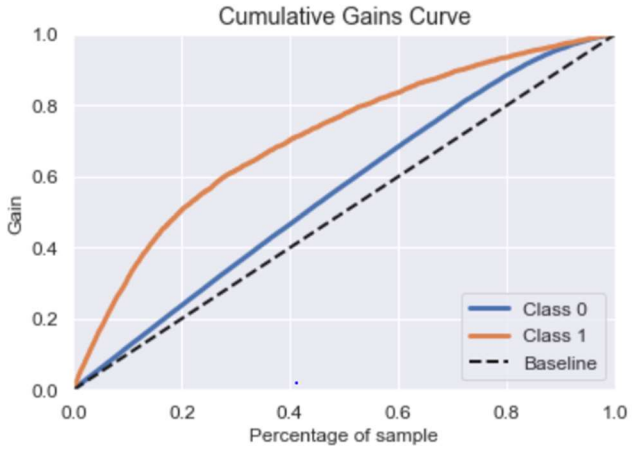
Discriminant Analysis模型在各項指標表現上均為最優秀。接著，我們進行更多資料處理和其他模型分析，嘗試得出更好的結果。

Fine-tune your models and combine them into a great solution

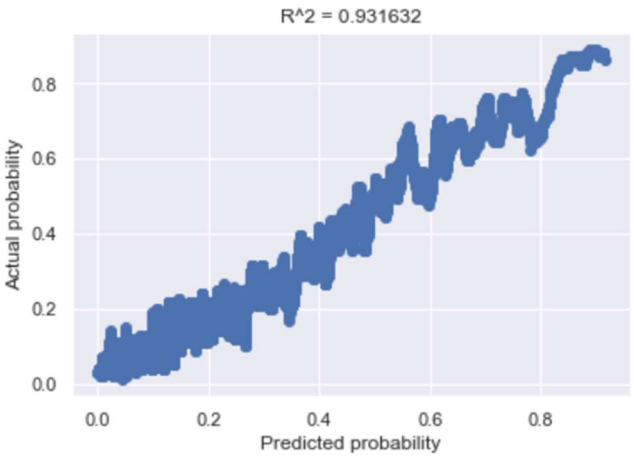
首先，我們試著嘗試了其餘的模型。我們首先採用Random Forest，並使用原始資料對其進行model training，也同時用前述的四項指標檢視預測準確度。

Random Forest

5) Cumulative Gain Curve

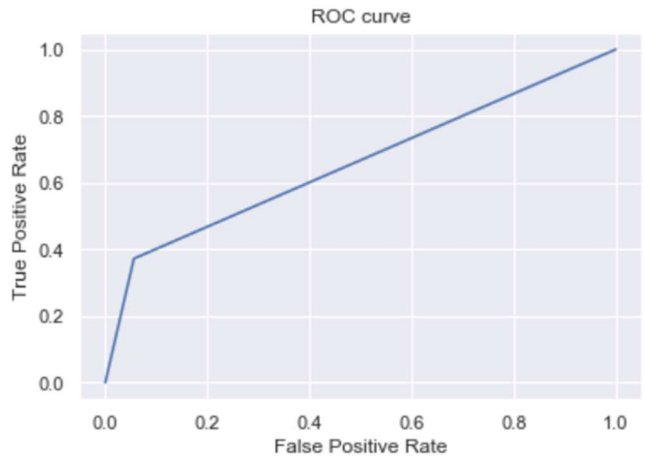


6) Sorting Smoothing Method(SSM)



7) ROC curve

```
0.6580035522006994
```



8) Error Rate

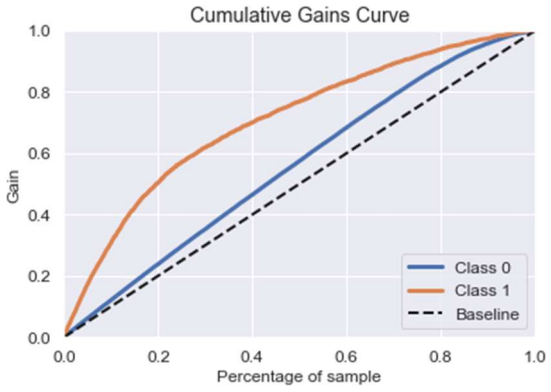
0.181

Random Forest模型雖然看似精準，但我們仍然希望能夠更加客觀地對資料進行分析。因此，我們首先對類別項的特徵進行生成虛擬變數(dummy variable)，接著再對所有資料進行標準化。除了Random Forest，我們還採用了SVM模型。

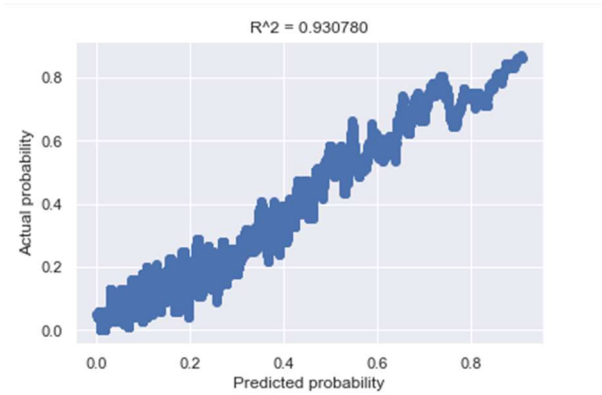
以下為其四項指標：

Random Forest With Data Processing

1) Cumulative Gain Curve

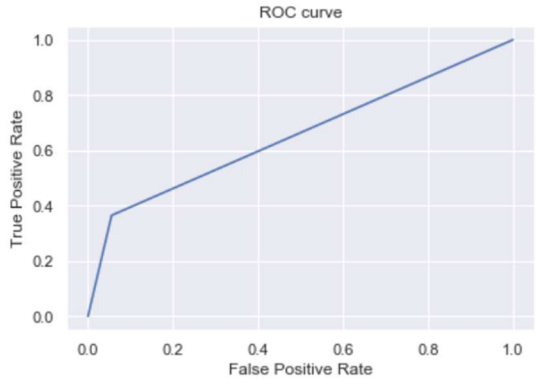


2) Sorting Smoothing Method(SSM)



3) ROC curve

```
0.65474029122871
```

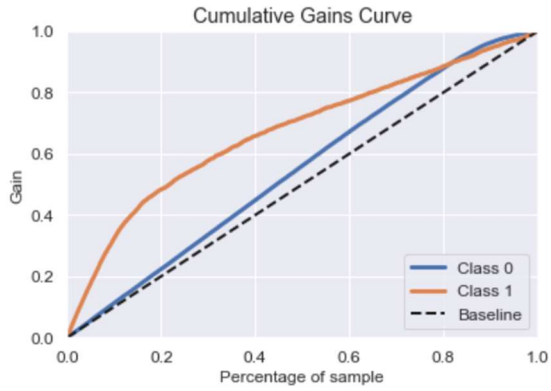


4) Error Rate

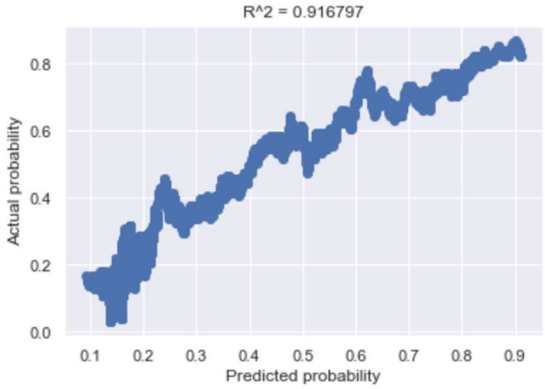
0.182

SVM With Data Processing

1) Cumulative Gain Curve

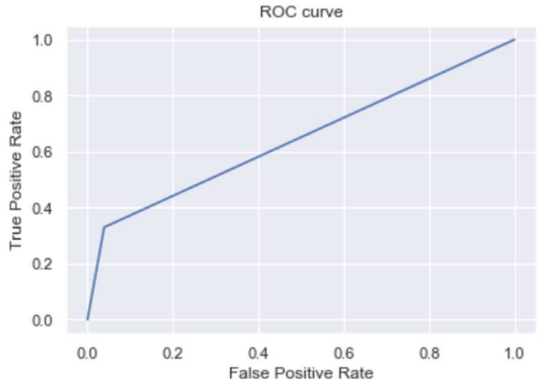


2) Sorting Smoothing Method(SSM)



3) ROC curve

```
0.6452604182098696
```



4) Error Rate

0.177

Present your solution

model	R^2	AOC	error rate
KNN	-2.634	0.549	0.246
Logistic	0.323	0.500	0.218
Discriminant Analysis	0.852	0.611	0.185
Naive Bayesian Classifier	-15.842	0.558	0.639
Decision Tree	-10.878	0.614	0.276
Random Forest	0.932	0.658	0.181
Random Forest With Data Processing	0.931	0.655	0.182
SVM With Data Processing	0.917	0.645	0.177

我們將改良模型過後的預測結果和其他預測結果的表格合併，可以發現採用新模型和我們對於資料的處理是有助於提升在此資料集的模型預測準確度的。

結論

綜觀以上八種狀況下的預測結果，可以明顯的看出不同狀況下的R^2有著相當大的差異，但error rate在除了Naive Bayesian Classifier外的情況下都相差不大，而AOC也差別不大，全距只有0.158。不過三種指標都可以明確的指出Random Forest With Data Processing是最好的預測模型，第二是SVM With Data Processing，雖然前者的error rate比後者大了0.005，但在AOC及R^2的表現都比較好。所以在此資料集，以信用卡使用者的個人資料來分析該使用者在下個月的拖欠機率，最佳的狀況是，在資料處理後使用Random Forest來預測。

參考論文:

The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients